# MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots

*Year: 2023 | Citations: 181 | Authors: Gelei Deng, Yi Liu, Yuekang Li*

---

## Abstract

Large Language Models (LLMs) have revolutionized Artificial Intelligence (AI) services due to their exceptional proficiency in understanding and generating human-like text. LLM chatbots, in particular, have seen widespread adoption, transforming human-machine interactions. However, these LLM chatbots are susceptible to"jailbreak"attacks, where malicious users manipulate prompts to elicit inappropriate or sensitive responses, contravening service policies. Despite existing attempts to mitigate such threats, our research reveals a substantial gap in our understanding of these vulnerabilities, largely due to the undisclosed defensive measures implemented by LLM service providers. In this paper, we present Jailbreaker, a comprehensive framework that offers an in-depth understanding of jailbreak attacks and countermeasures. Our work makes a dual contribution. First, we propose an innovative methodology inspired by time-based SQL injection techniques to reverse-engineer the defensive strategies of prominent LLM chatbots, such as ChatGPT, Bard, and Bing Chat. This time-sensitive approach uncovers intricate details about these services' defenses, facilitating a proof-of-concept attack that successfully bypasses their mechanisms. Second, we introduce an automatic generation method for jailbreak prompts. Leveraging a fine-tuned LLM, we validate the potential of automated jailbreak generation across various commercial LLM chatbots. Our method achieves a promising average success rate of 21.58%, significantly outperforming the effectiveness of existing techniques. We have responsibly disclosed our findings to the concerned service providers, underscoring the urgent need for more robust defenses. Jailbreaker thus marks a significant step towards understanding and mitigating jailbreak threats in the realm of LLM chatbots.